Ann Oper Res (2017) 252:63–84
DOI 10.1007/s10479-015-2058-3

Staffing of time-varying queues using a geometric discrete time modelling approach

Xi Chen[1] · **Dave Worthington**[1]

Published online: 23 November 2015

Abstract Many queueing systems with time-dependent arrivals require time-dependent staffing to provide satisfactory service levels at reasonable cost. Feldman et al. (Manag Sci 54(2):324–338, 2008) proposed an iterative staffing algorithm designed to deliver time-stable performance in which successive iterations were evaluated via simulation experiments. In this paper we present and evaluate an analytical queueing model combined with an iterative staffing algorithm to be used for setting staffing levels to achieve time-stable performance in call centre type queues. Empirical results show that the method to be considerably faster than simulation based methods and considerably more accurate than the industry standard analytical methods.

Keywords Time-dependent analysis · Multi-server queues · Discrete time modelling · Call centre staffing

1 Introduction

In this paper we develop and evaluate analytical methods to determine appropriate staffing levels in call centres and in other multi-server queueing systems with time-dependent arrival rates. The importance of time-dependent as opposed to steady-state analysis for many real queueing systems is well established. Call centres (see for example Gans et al. 2003), communication networks (see for example Abdalla and Boucherie 2002), healthcare (see for example Izady and Worthington 2012; Bekker and de Bruin 2010) and traffic flows (see for example Griffiths et al. 1991) are all areas where time-dependent analysis can be essential.

In order to set staffing requirements in a time-varying arrival rate situation, it is common practice for call centre workforce management to use stationary models in a nonstationary manner—that is, to chop time into segments and then use a stationary model for each segment

✉ Dave Worthington
d.worthington@lancaster.ac.uk

[1] Lancaster University, Lancaster, UK

(Green et al. 2007). The *pointwise stationary approximation* (PSA) is a popular approximation method of this type (Green and Kolesar 1991). PSA gives a time-dependent description of performance based on a stationary model. It uses the instantaneous arrival rate that prevails at each moment in time to describe the performance at that time, and it assumes that steady state is instantly achieved at each moment. Thus the PSA is usually applied with the well-known steady-state *Erlang-C* or *Erlang-A* formulae (which assume systems of the form $M/M/s$ and $M/M/s+M$ respectively) to determine the piecewise staffing requirements to achieve target service levels. Green et al. (2007) showed that PSA-based approaches produce good approximations when service times are short (e.g. 3 min on average) and the quality-of-service standard is high, e.g. 90 % of the calls are answered immediately. Under such circumstances there is less likely to be a queue, and therefore steady state can be achieved faster.

If service levels are high (i.e. systems that are *quality driven*) but service times are not short, Green et al. (2007) describe how PSA-based approaches need to be modified to take account of the time lag that can occur between peaks in arrival rates and peaks in congestion. In such circumstances the *modified-offered-load* (MOL) based staffing algorithm approximation, introduced by Jennings et al. (1996), can be used. Feldman et al. (2008) pointed out that in many real applications, the MOL approximation can be well approximated itself by lagged PSA, i.e. adjusting the instantaneous arrival rate by a time shift equal to the mean service time.

However, none of these methods can be expected to work well in *efficiency driven* systems where service level targets do not require low probabilities of delay, in which case the lags between peaks in arrivals and peaks in congestion will substantially exceed those indicated by the MOL approach. In addition these approaches also fail when the traffic intensity exceeds one, which is not unusual in real call centres, where forecast errors in the time-varying arrival rates can often lead to insufficient numbers of agents (Chassioti and Worthington 2004).

In response to these circumstances Feldman et al. (2008) proposed a simulation-based iterative staffing algorithm (ISA) to achieve time-stable performance for systems of the form $M_t/G/s_t+G$. They were also able to prove convergence to the desired result under certain additional assumptions, and that it converged in practice for a wide range of empirical cases. Two limitations of this approach were that it only considered very short staffing intervals, and that system performance was limited to the delay probability, i.e. target service levels of the form $P(wait>0)\leq\alpha$. Defraeye and Van Nieuwenhuyse (2013) have since further developed the approach to incorporate longer staffing intervals and target service levels of the form $P(wait>\tau)\leq\alpha$, as well as improving the algorithm's stopping rule for small systems.

An important feature of these approaches is that they seek to achieve time-stable performance targets even though time-stable target service levels are rare in practice, not least because of measurement problems. However time-stable performance is often a desirable quality, and staffing levels that provide time-stable performance will often provide a good starting point around which to fine tune actual staff rosters.

Despite the clear potential of simulation-based methods, when describing call centre staffing methods in practice, Koole (2013) comments that an analytical approach, namely the PSA, is the industry standard. The preference for the PSA is easy to understand, as simulation can lead to very long execution times to search for the best staffing solutions (Koole 2013). In this paper we present an analytical queue modelling approach which offers substantial computational savings in comparison to a simulation-based approach, and substantial improvements in accuracy when compared to the PSA and lagged PSA approaches.

The basis of this approach is discrete time modelling (DTM), the basic principles of which can be traced to Galliher and Wheeler (1958) who studied $M(t)/D/c$ systems. Dafermos and Neuts (1971) suggested use of the DTM approach to approximate continuous time queues,

while Neuts (1973) suggested using numerical methods to solve the associated recurrence relations. More recent work has developed DTM for non-stationary multi-server queues. Worthington and Wall (1999) provide an overview of much of this work with respect to queue length behaviour, and Wall and Worthington (2007) extend the approach to model waiting time behaviour. In this paper we combine DTM, rather than simulation modelling, with an iterative staffing algorithm to achieve time-stable performance.

Although DTM is an analytical approach, its numerical solution can still cause computational issues if the statespace of the underlying discrete time Markov process is large. Initial investigations using DTM showed that this would be problematic in the context of the repeated runs needed in an iterative staffing method. In many respects the statespace of traditional DTM is similar to that used when applying the method of phases to model $M/E_k/S$ systems, i.e. the state of the system needs to record not only the number in the system but also the numbers of customers with 1, 2, 3,. . .. phases of service remaining. In DTM the state of the system needs to record the number in the system and the numbers of customers with 1, 2, 3,. . .. units of service remaining (Worthington and Wall 1999). Hence, as in continuous time queues, the statespace is dramatically reduced if the service time is assumed to be memoryless, i.e. Exponential in continuous time or Geometric in discrete time.

Although one of the strengths of the traditional DTM is that discrete service time distributions could always be chosen to match the mean and squared coefficient of variation (SCV) of any real service time distribution, empirical work with call centre based scenarios without abandonments and with state-dependent balking, Chassioti and Worthington (2004) and Chassioti et al. (2014) showed that matching the mean service time (and hence the traffic intensity) was much more important than matching the SCV of service time, particularly in systems in which arrival rates vary quite quickly. Hence there was reason to believe that a 'Geometric DTM' in which service times were constrained to taking a Geometric distribution might achieve sufficient accuracy for the purposes of setting staffing levels in a call centre setting.

On the other hand Whitt (2005) presented empirical results for the steady-state behavior of large call centres where the call centre performance was more sensitive to the distribution of abandonment times than to the distribution of service times, and in particular to the shape of the left-hand tail of the distribution. Hence it is also important to investigate the impact of different abandonment time scenarios on the effectiveness of the 'Geometric DTM'.

The purpose of this paper is therefore to present and evaluate a Geometric DTM approach combined with an iterative staffing algorithm to be used for setting staffing levels to achieve time-stable performance in call centre type queues of the form $M_t/G/s_t$, both with and without abandonments. The Geometric DTM is introduced next in Sect. 2, and the iterative staffing algorithm used in this work is described in Sect. 3. The qualities of the solutions are then evaluated using a simulation framework in Sect. 4, and in particular they are shown to be considerably better than the industry standard PSA-based methods and considerably faster than simulation based methods in many circumstances. Finally in Sect. 5 the main conclusions to this work are highlighted and discussed, as well as ideas for further work.

2 Geometric DTM models

In order to reduce the computation requirements of the traditional DTM approach for staffing purposes, we devise a DTM algorithm with Geometric distribution of service times. The Geometric DTM is a simplified version of the traditional DTM algorithm because of the

memoryless property of the Geometric distribution. This implies that, in the discrete time context, if there is a customer in service at the current epoch t, the probability of the service completion in the next time epoch $t+1$ will be a constant factor g, and is independent of the service time so far. As a result, the statespace associated with the Geometric DTM only needs to record the number of customers in the system, greatly reducing the memory and computational requirements of the method in comparison to the traditional DTM approach.

In this section we first introduce a basic Geometric model ($M_t/Geom/s$), we then extend it to incude Geometric abandonments ($M_t/Geom/s+Geom$) and finally extend it further to include time-dependent staffing levels ($M_t/Geom/s_t+Geom$).

2.1 $M_t/Geom/s$

The basic Geometric DTM models the queue as a Markov chain, where we denote the state at time t as n_t—the number of customers in the system at time t, where n_t can take the values 0, 1, 2,. . ., L. We make the following assumptions:

- The time of operation of the system is divided into a set of equal non-overlapping intervals, often referred to as slots. The epochs of each slot are labelled by the integers t $= 0, 1, 2 \ldots$, where 0 is the beginning of the operation and the length of each interval represents one unit of discrete time. The system is only observed at each epoch.
- The arrival process is random at rate λ(t) between time t and t$+1$. The probability distribution of the number of arrivals between two adjacent epochs is therefore Poisson with mean λ(t) and is independent of arrivals in other slots. The arrivals are assumed to enter the system at the end of the slot in which they arrive.
- There are s servers in the system.
- There is an upper limit on the numbers allowed in the system—L; any arrivals when the system is full are assumed to be lost. As arrivals occur at the ends of slots, any losses are also assumed to happen at the ends of slots.
- The Geometric DTM service times have a Geometric distribution with ratio g, i.e. the mean service time is $\frac{1}{g}$, variance is $\frac{(1-g)}{g^2}$, and hence its SCV is $(1-g)$ For each service, the probability of service completion in the current time slot is always g.
- The arrival and service processes are independent—so their joint probabilities between any epochs are the products of their separate probabilities.

These assumptions mean that the $M_t/Geom/S$ queueing system queues can be formulated as a time-inhomogeneous Markov chain, and hence the full time-dependent distribution of the number in the system can be evaluated using the simple relationship:

$$\boldsymbol{\pi}\,(t+1) = \boldsymbol{\pi}\,(t)\,\boldsymbol{P}\,(t) \quad \text{for } t = 0, 1, 2, \ldots\ldots\ldots$$

where $\boldsymbol{\pi}(t)$ is the vector of state probabilities at time t, i.e.

$$\pi_n\,(t) = Pr\,(n_t = n)\,;$$

and $\boldsymbol{P}(t)$ is the matrix of transition probabilities for the time interval $(t, t+1]$. The transition probabilities are obtained by considering the events that can occur from epoch t to epoch $t+1$, as follows.

Suppose the number in the system at time t is n_t, then the total number of customers in the system between t and $t+1$ is n_t, hence the number of customers in service between t and $t+1$ is

$$c_t = min\,(n_t, s) \tag{1}$$

Let d denote the number of departures (i.e. service completions) between epoch t and $t+1$, and let $S_{d|c}(t)$ denote the probability of d service completions between t and $t+1$ conditional on c customers in service at time t. As completions are independent with probability g, the number of completions $\sim$ Binomial (c, g), and we have

$$S_{d|c}(t) = \frac{c!}{d!\,(c-d)!} g^d (1-g)^{(c-d)} \tag{2}$$

(Note that the number of customers in service at time t (c_t) will depend on t, but we omit the subscript t here for ease of notation.)

We can also calculate the possible numbers of arrivals at the end of the interval $(t, t+1]$, and their probabilities. Let r = number of arrivals at the end of the interval $(t, t+1]$, $r = 0, 1, 2 \ldots L$, with probability:

$$V_r(t) = \begin{cases} \frac{e^{-\lambda(t)}\lambda(t)^r}{r!}, & r < L \\ 1 - \sum_{i=0}^{L} \frac{e^{-\lambda(t)}\lambda(t)^r}{r!}, & r = L \end{cases} \tag{3}$$

as arrivals in $(t, t+1]$ are random at rate $\lambda(t)$. Hence the number of customers in the system at epoch $t+1$ is n_{t+1}, where:

$$n_{t+1} = min\,(L, n_t - d + r)\,. \tag{4}$$

As with the original DTM, these relationships can be easily implemented using a forward recurrence algorithm. Each possible state at epoch t is considered in turn, and in each case all possible events between epoch t and epoch $t+1$ are considered and their probabilities are calculated together with the resulting state at epoch $t+1$. Probabilities of the different resulting states are then accumulated to give the state probabilities at epoch $t+1$.

The algorithm starts at epoch 0 with a starting condition; for example, the system starts empty, i.e. $\pi(0) = (1, 0, 0, \ldots)$. In this way, the probability distribution of the queueing system's states at each epoch of the whole time period T can be computed. See "Appendix 1" for details of the algorithm.

2.2 $M_t/Geom/s{+}Geom$

The geometric DTM can be extended to model call abandonment behaviour where we assume the time-to-abandon has a Geometric distribution with mean $1/f$, i.e. the probability that a customer not in service at time t abandons by the next time epoch $t+1$ is the constant factor f, and is independent of the queueing time so far, and of the arrival and departure processes. The number of customers not in service is:

$$q_t = min\,(0, n_t - s) \tag{5}$$

In this model, the possible number of abandonments and their associated probabilities can be generated in a similar way to the number of service departures in the previous model. Let a denote the number of abandonments in $(t, t+1]$, and let $A_{a|q}(t)$ denote the probability of a abandonments between time t and $t+1$, conditional on q customers in the queue at time t.

As abandonments are independent with probability f, the number of abandonments $\sim$ Binomial (q, f), and we have

$$A_{a|q}(t) = \frac{q!}{a!\,(q-a)!} f^a (1-f)^{(q-a)} \tag{6}$$

(Note that the number of customers not in service at time t (q_t) will depend on t, but we omit the subscript t here for ease of notation.)

Hence the number of customers in the system at epoch $t+1$ is n_{t+1}, where:

$$n_{t+1} = min\,(L, n_t - d + r - a)\,.$$

The algorithm to implement the associated forwards recurrence relationships is provided in "Appendix 2".

2.3 $M_t/Geom/s_t+Geom$

In this model the number of staff change at specified epochs, according to a staffing plan. At such epochs we define:

- The number of servers at time t: s_t
- The number of servers after the planned staffing change:s_{t^+}

When extra servers are scheduled to join the system at time t, i.e. $s_{t^+} > s_t$, this can be incorporated very simply into the previous formulations and algorithms by simply updating c_t to:

$$c_t = min\,(n_t, s_{t^+}) \tag{7}$$

When some servers are scheduled to leave the system, i.e. $s_{t^+} < s_t$, it is important to know whether the server departure policy is pre-emptive or exhaustive. A pre-emptive policy assumes that when the servers are scheduled to leave, the service in progress with those servers will be interrupted and the customers at those service points rejoin the queue. In contrast, an exhaustive policy is defined as the case where, when the servers are scheduled to leave at the end of their shift, any services in progress with those servers will have to be completed before they leave (Ingolfsson 2005). Compared to the pre-emptive policy, the exhaustive policy is widely considered as a more realistic case in real call centre queueing service systems (Ingolfsson 2005).

Under the pre-emptive policy any customers who lose their servers partway through service at time t will cause a reduction in c_t, a corresponding increase in q_t and no change in n_t. As a consequence in the interval $(t, t+1]$ c_t customers will be subject to service completions with probability g and q_t customers will be subject to abandonments with probability f, exactly as required under the pre-emptive policy. Hence the previous forward recurrence algorithms will implement this policy automatically.

For the exhaustive policy in consideration, if the number of customers in the system $n_t \leq s_{t^+}$, the change in number of servers has no impact on the state of the system at time t (here we assume that leaving servers are not busy, or have been able to hand their customer over to a server who is not busy), and hence the existing forward recurrence relationships deal with this case automatically.

However when $n_t > s_{t^+}$, we need to consider the number of customers in the system n_t in two parts: n_t^a and n_t^b.

(i) Let $n_t^a = \min(s_t, n_t) - s_{t^+}$: the number of customers being served by the servers who are about to end their shifts. These customers do not have any impact on the waiting time of future customers as they are served by the leaving servers.
(ii) Let $n_t^b = n_t - n_t^a$: the rest of customers in the system.

The part (i) customers can be considered as a completion-only queueing system, as these n_t^a customers simply remain in the system until their services are completed. Hence, (omitting the subscript t here for ease of notation as before) if d^a is the number of service completions

in $(t, t+1]$, then d^a has a Binomial distribution i.e. the probability of d^a completions conditional on n^a customers in service with servers about to leave ($S_{d^a|n^a}(t)$) is given by:

$$S_{d^a|n^a}(t) = Pr\{d^a|n^a\} = \frac{n^a!}{d^a!\,(n^a - d^a)!} g^{n^a} (1-g)^{(n^a - d^a)} \tag{8}$$

Therefore, for this completion-only system:

$$n^a_{t+1} = min\left(0, n^a_t - d^a\right) \tag{9}$$

which is easily incorporated into the forward recurrence relationships in the same way as before.

The part (ii) customers on the other hand now behave exactly as in the pre-emptive policy system until the next change of shift, and hence can be modelled using the previous forward recurrence relations until that point, when consideration again needs to be given to whether the change in number of servers results in any more completion-only customers.

3 Staffing algorithm

Staffing periods (i.e. periods for which the staffing level remains unchanged) are defined as multiples of the basic timestep and hence take the form $(t, t+m]$ for m $\geq$ 1. In order to achieve time-stable performance throughout the day we require that the performance target is achieved for each staffing period through the period of operation.

Unlike a simulation-based staffing algorithm, an algorithm based on Geometric DTM does not require multiple runs to provide estimates of system performance under any particular staffing pattern. Furthermore the forward recurrence method of evaluation in Geometric DTM means that when deciding the staffing levels up to time t it is only necessary to run the forward recurrence calculations up to time t if the performance measure is delay probability. If the performance measure relates to the probability that a customer's delay (i.e. queueing time) exceeds u, the forward recurrence equations only need to be evaluated up to time $t+u$ to calculate the probability that customers arriving by time t enter service by time $t+u$.

For each staffing period the search for the appropriate staffing level is iterative, and is based on the iterative staffing algorithm described in Feldman et al. (2008). Starting with a very large number of servers (e.g. $s_1 = L$), Geometric DTM is used to calculate the time-dependent distributions of customers in the system for the staffing period $\{\pi_n^{(1)}(\tau)$ for $\tau : t$ to $t+m$; $n : 0$ to $L\}$, and the associated conditional probabilities that customers achieve the delay time target u, $\{p^{(1)}(delay < u|finds\, n)$ for $n : 0$ to $L\}$. s_2 is then found by choosing s_2 to be the smallest n such that the target service level would be achieved even if everybody who arrived to find more than n in the system failed to meet the target waiting time, i.e.:

$$s_2 = \arg\min\left\{k : \left(\sum\nolimits_{\tau=t}^{t+m} \sum\nolimits_{n=0}^{k} [\pi_n^{(1)}(\tau) \times p^{(1)}(delay < u|finds\, n\right) > \alpha\right\}$$

Because s_1 is very large, this initial run of Geometric DTM provides a large underestimate of the congestion levels that would occur under the desired staffing level, and hence in practice s_2 is an underestimate of (or possibly equal to) the desired staffing level. The process is then repeated using Geometric DTM to calculate the time-dependent distributions of customers in the system for the staffing period with s_2 servers to give a new estimate s_3 of the desired staffing level. Because s_2 is an underestimate of (or possibly equal to) the desired staffing level, s_3 is an overestimate of (or possibly equal to) the desired staffing level. Hence s_2 and

s_3 provide finite bounds, which means that a binary search has no convergence problems, and very quickly gives the minimum s such that the target service level is achieved for the staffing period.

We note that Feldman et al. (2008) were able to prove that their algorithm converged to the desired staffing levels for systems of the form $M_t/M/s_t+M$ (under their assumption of very short staffing intervals), and their empirical results showed evidence that it worked on a wider set of systems. Our experience of using their iterative approach on cases which include systems of the form $M_t/G/s_t$ and do not assume very short staffing intervals, support the contention that it is robust and widely applicable. We also found that the introduction of the binary search, once the initial bounds s_2 and s_3 had been found, worked well and was often quicker than the original algorithm.

4 Results

To investigate how the Geometric DTM-based ISA algorithm (Geo-DTM+ISA) performs, in this section we apply the algorithm to a range of realistic call centre test cases.

The test cases are based on data from a medium sized insurance service call centre in the UK during the year 2004–2005. The call centre operates for 10 h a day, with the typical time-varying arrival rate across the business hours. The average half-hourly call volumes used are shown in Fig. 1, and the mean service time $E(S) = 247$ s.

For each of the test cases we first use Geo-DTM+ISA to recommend staffing levels, and then perform multiple discrete event simulation runs of the system of interest with the recommended staffing to investigate whether the target service levels are achieved. Early test runs showed that 2000 simulation runs were necessary to obtain results that were accurate to $+/-2\,\%$ with 95 % confidence, and hence 2000 runs were used in all test cases.

The results in this section are presented with the following five aims in mind.

1. To investigate whether the discretisation of a continuous system and the staffing algorithm introduce any significant sources of error. This is achieved by testing Geo-DTM+ISA on systems of the type $M_t/M/s_t$ and $M_t/M/s_t+M$, see Sects. 4.1 and 4.2. This is the case where the approach is most likely to be successful as the memoryless continuous service time is being approximated by the memoryless discrete distribution.

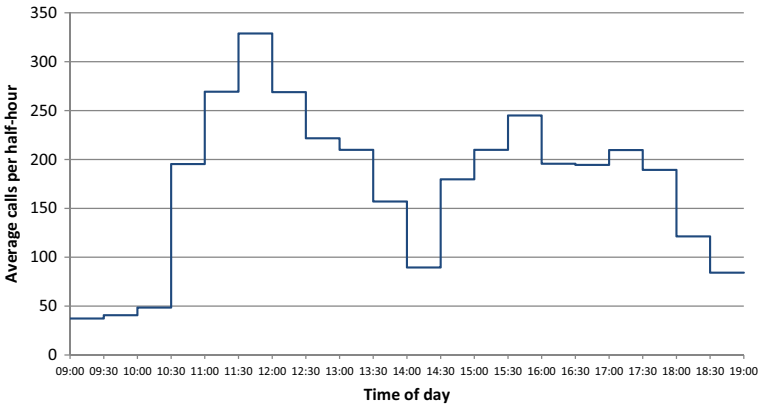

Fig. 1 Half-hourly average call volumes to a medium sized insurance service call centre

2. To investigate the improvements that Geo-DTM+ISA offers over the industry standard PSA method, and the lagged PSA method. Both these methods assume Exponential service times, and so empirical results for the $M_t/M/s_t$ and $M_t/M/s_t+M$ cases provide a fair comparison, see Sects. 4.1 and 4.2.
3. To investigate the extent to which the performance of Geo-DTM+ISA deteriorates when used on cases where the service time is not Exponential. This is achieved by testing Geo-DTM+ISA on systems of the type $M_t/G/s_t$, where the service times take a range of Lognormal and Beta distributions with squared coefficients of variation (SCVs) ranging from 0.077 to 2.0. See Sect. 4.3.
4. To investigate the extent to which the performance of Geo-DTM+ISA deteriorates when used on cases where the abandonment time is not Exponential. This is achieved by testing Geo-DTM+ISA on systems of the type $M_t/G/s_t+$G, where the abandonment times take a range of Lognormal and Erlang distributions, with squared coefficients of variation (SCVs) ranging from 0.5 to 4.0. See Sect. 4.4.
5. To provide a rough comparison of the computational performance of Geo-DTM+ISA in comparison to a simulation based method, see Sect. 4.5.

4.1 $M_t/M/s_t$

Figure 2 compares the system performance of the Geo-DTM+ISA, the Erlang C formula combined with PSA (Erlang-C+PSA) and the Erlang C formula combined with lagged PSA (Erlang-C+LPSA), for systems of the form $M_t/M/s_t$, all under the pre-emptive policy. Four different time service factors (TSF) were considered, 80, 60, 40 and 20 % served within 0 s. The results obtained for the intermediate service levels (i.e. 60 and 40 %) showed characteristics entirely in line with those observed for the more extreme service levels (see Chen 2014 for details), and hence are omitted from the results presented in this paper for clarity of presentation. Other TSFs of the form 'x % served within τ seconds' were also investigated and produced results of a very similar nature. The output performances presented are the actual service levels of the simulations using the staffing patterns provided by the different staffing methods.

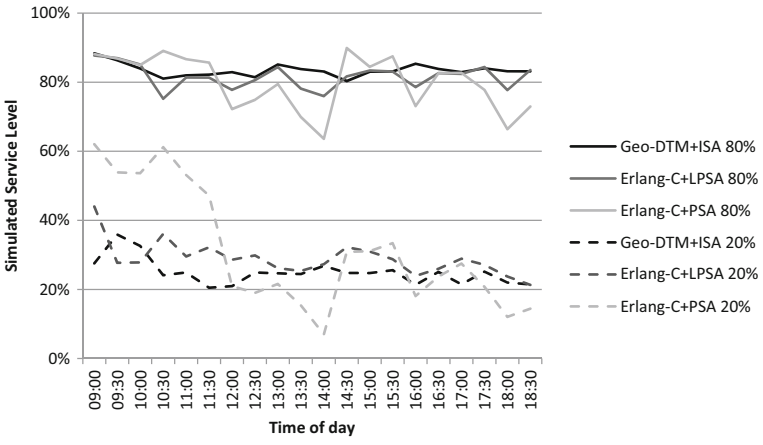

Fig. 2 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA, Erlang-C+PSA and Erlang-C+LPSA methods (TSF within 0 s, pre-emptive policy)

It is evident from Fig. 2 that Geo-DTM+ISA consistently achieves service levels which are stable and adhere to the specified service target lines throughout the day. On the other hand the service levels produced by the Erlang-C+PSA and LPSA methods fluctuate over time and the service targets are sometimes not achieved and sometimes over-achieved (i.e. provide more staff than needed). For example, if we look at the TSF = 20 % lines at 11:00 specifically, Geo-DTM+ISA achieves a service level very close to target at 22 % but Erlang-C+PSA is very over-staffed with a service level at over 50 %. Erlang-C+LPSA is also overstaffed and achieves a service level of around 30 %. If we look at the same TSF at 14:00, we can see that Geo-DTM+ISA still achieves close to target service level at 25 %, but Erlang-C+PSA now achieves too low a service level at <10 %. On the other hand, at 14:30, Erlang-C+LPSA is overstaffed at over 30 %, while Geo-DTM+ISA is at 23 %.

The relative performance of Erlang-C+PSA and Erlang-C+LPSA is as expected with the lagged model generally performing better than the unlagged model. However, Erlang-C+LPSA still tends to overestimate the staff requirement throughout the day at the low TSFs and underestimate the requirement at the high TSFs. The impacts on the staffing levels of the three methods are shown in Figs. 3 and 4. In general the Erlang-C+LPSA is closer to Geo-DTM+ISA than is Erlang-C+PSA, and the differences between the three methods are bigger for the lower TSF (Fig. 4) than for the higher TSF (Fig. 3).

These observations are typical of more extensive tests which considered TSFs expressed in terms of customers starting service within 20 s (rather than 0 s) and an exhaustive policy rather than a pre-emptive policy, see Chen (2014) for details. For example Fig. 5 compares the system performance of the Geo-DTM+ISA against Erlang-C+PSA and Erlang-C+LPSA under the exhaustive policy, and TSFs of 80 and 20 % served within 20 s. Geo-DTM+ISA continues to work very well, and Erlang-C+PSA and Erlang-C+LPSA continue to show the same weaknesses. Overall our results indicate that Geo-DTM+ISA shows similar advantages over Erlang-C+PSA and Erlang-C+LPSA for both pre-emptive and exhaustive policies, as it mimics both policies quite accurately. Performance relative to the PSA-based methods

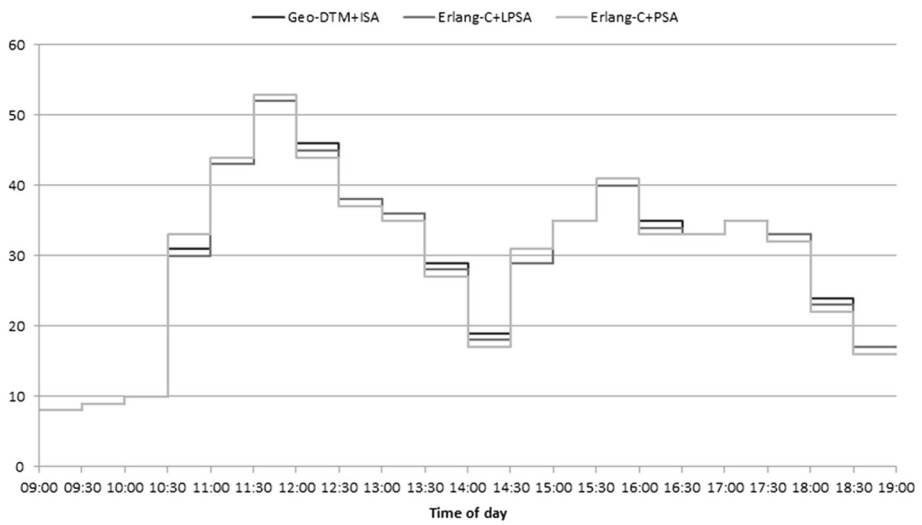

Fig. 3 Staffing levels required according to Geo-DTM+ISA, Erlang-C+PSA and Erlang-C+LPSA methods (TSF 80 % within 0 s, pre-emptive policy)

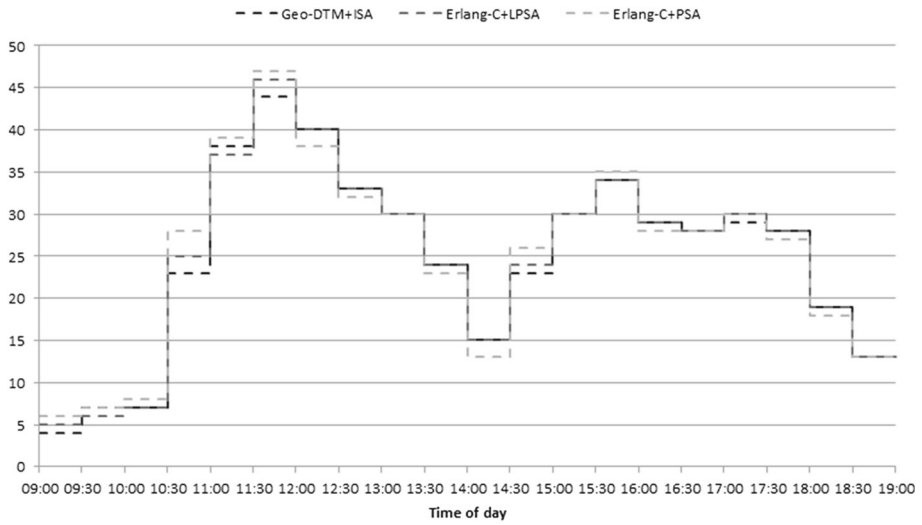

Fig. 4 Staffing levels required according to Geo-DTM+ISA, Erlang-C+PSA and Erlang-C+LPSA methods (TSF 20 % within 0 s, pre-emptive policy)

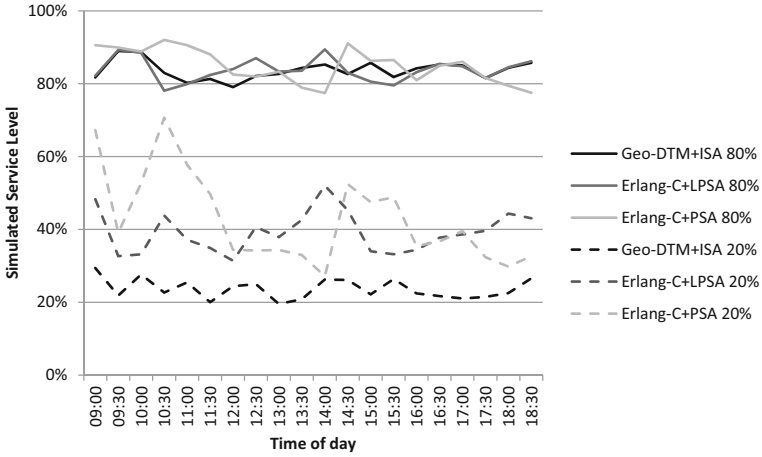

Fig. 5 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA, Erlang-C+PSA and Erlang-C+LPSA methods (TSF within 20 s, exhaustive policy)

therefore depends mainly on the performance of the latter, which can be quite variable between the the two policies.

Also notable from Figs. 2 and 5 is that Erlang-C+PSA and Erlang-C+LPSA perform better for systems in which queues rarely occur, i.e. those with the higher TSF values. This is because these are systems with relatively low traffic intensities, which therefore tend to settle to steady state more quickly, and hence the PSA assumptions are closer to being true (Green et al. 2007).

Very similar sorts of results were also obtained for smaller call centres. For example, Fig. 6 is comparable with Figure 2, but for a call centre with arrival rates multiplied by a factor 0.25.

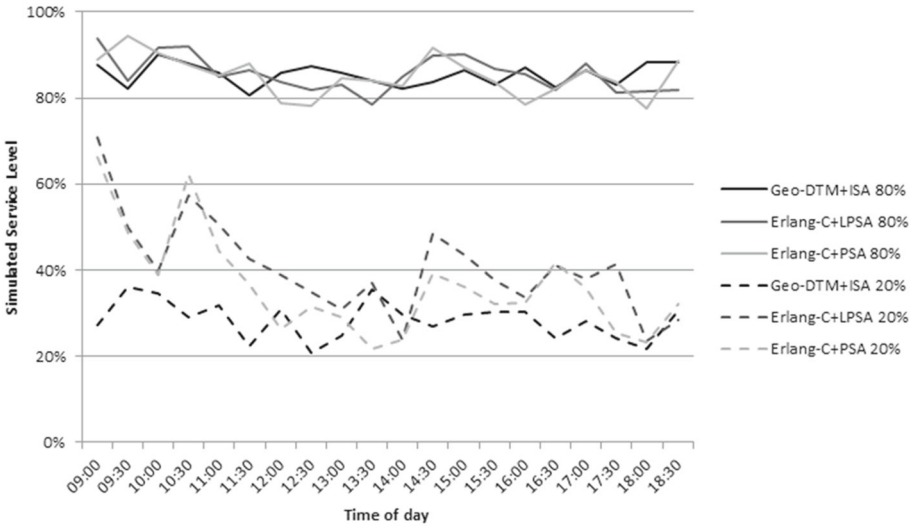

Fig. 6 Smaller Call Centres: Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA, Erlang-C+PSA and Erlang-C+LPSA methods (TSF within 0 s, pre-emptive policy)

Geo-DTM+ISA consistently achieves the TSFs, whilst Erlang-C+PSA and Erlang-C+LPSA produce less stable and less satisfactory staffing patterns, particularly for lower TSFs.

4.2 $M_t/M/s_t+M$

Similar investigations were undertaken to evaluate the performance of Geo-DTM+ISA for systems with abandonments, and to again compare the performance with the industry standard method, i.e. Erlang A formula combined with PSA (Erlang-A+PSA) and with the Erlang A formula combined with lagged PSA (Erlang-A+LPSA). In addition to the issues of service level and pre-emptive/exhaustive policy already observed, there is the additional issue of customer 'patience' in systems with abandonments. Hence test cases investigated service levels and pre-emptive/exhaustive policies as before, and in addition experimented with mean abandonment times ranging from 0.5 × mean service time through to 2 × mean service time.

Figure 7 shows the service level performance of Geo-DTM+ISA and the two benchmark methods, Erlang-A+PSA and Erlang-A+LPSA respectively. All staffing methods are aiming for two TSFs (20 % within 0 s and 80 % within 0 s) under the pre-emptive policy. The mean time to abandon is double the mean service time. As noted for the $M_t/M/s_t$, Geo-DTM+ISA again consistently achieves service levels which are stable and adhere to the specified service target lines throughout the day. On the other hand the service levels produced by the Erlang-A+PSA and LPSA methods fluctuate over time and the service targets are sometimes not achieved and sometimes over-achieved.

As for the $M_t/M/s_t$ system, these results are typical of more extensive tests which considered target service expressed in terms of customers served with 20 s (rather than 0 s) and an exhaustive policy rather than a pre-emptive policy, see Chen (2014) for details. Geo-DTM+ISA continues to perform well when stronger abandonments are introduced, although as customers become less patient the weaknesses of Erlang-A+PSA and Erlang-A+LPSA become less important. This seems to be because, as noted by Chassioti et al. (2014), aban-

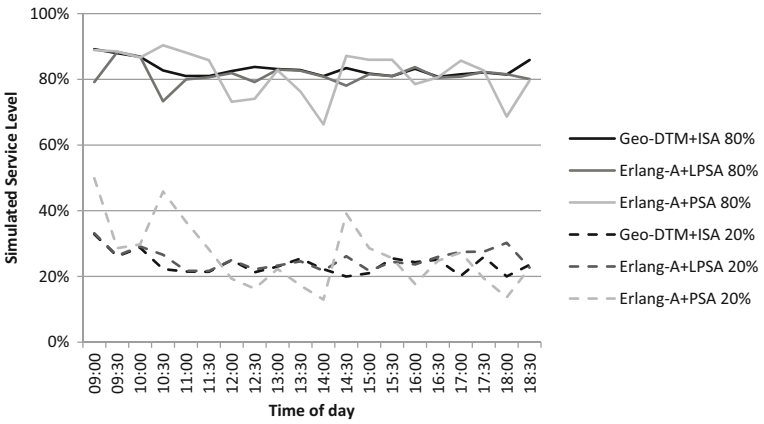

Fig. 7 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA, Erlang-A+PSA and Erlang-A+LPSA methods (TSF within 0 s, pre-emptive policy, with time to abandon = 2 × service time)

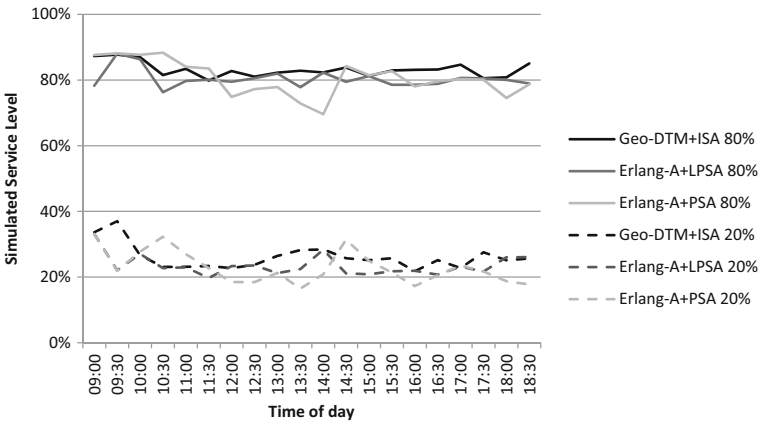

Fig. 8 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA, Erlang-A+PSA methods and Erlang-A+LPSA methods (TSF within 20 s, pre-emptive policy, with time to abandon = 0.5 × service time)

donments increase the speed with which systems settle to steady state, and hence the PSA assumptions are closer to being true. For example Fig. 8 is for identical settings as Fig. 7, except that mean time to abandon = 0.5 × mean service time, and whilst Geo-DTM+ISA continues to perform very well, Erlang-A+PSA and Erlang-A+LPSA are now also performing reasonably well.

4.3 $M_t/G/s_t$

To investigate the extent to which the performance of Geo-DTM+ISA deteriorates when used on cases with non-exponential service times, it has been tested on systems of the type $M_t/G/s_t$, where the service times take a range of Lognormal and Beta distributions with squared coefficients of variation (SCVs) ranging from 0.077 to 2.0. (Note that the Exponential distribution has SCV = 1.0). Figure 9 shows results for a Beta[2,5] distribution of service time

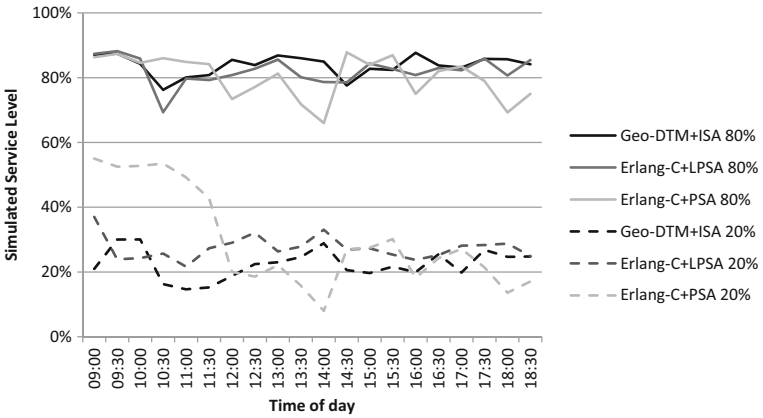

Fig. 9 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA, Erlang-C+PSA and Erlang-C+LPSA methods (Beta [2,5] service, TSF within 0 s, pre-emptive policy)

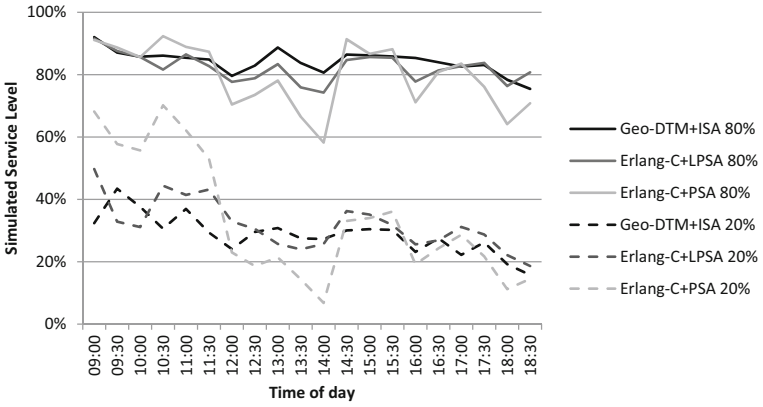

Fig. 10 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA, Erlang-C+PSA and Erlang-C+LPSA methods (Lognormal (SCV = 2.0) service, TSF within 0 s, pre-emptive policy)

(SCV = 0.313) and Fig. 10 shows results for a Lognormal with an SCV = 2.0. These results are typical of the range of results obtained (see Chen 2014 for details). In particular when SCV $<$ 1.0 or SCV $>$ 1.0, Geo-DTM+ISA sometimes deviates below the target service level (when SCV $<$ 1.0), and sometimes deviates below or unnecessarily above it (when SCV $>$ 1.0). In our results it continued to outperform Erlang-C+PSA and Erlang-C+LPSA.

These findings are not unexpected as an important feature of the traditional DTM was its matching of the service time in mean and SCV. In the previous sections the memoryless Geometric distribution has been used to approximate exponential distributions. As a Geometric distribution with mean $1/g$ has a SCV of $(1-g)$, by choosing a very small step size such that the mean service time was 500 steps, the Geometric distribution had an SCV = 0.998. Clearly by varying the step size a Geometric distribution can be found to have any SCV $<$1.0, and this does not cause any of the computational challenges associated with the traditional DTM.

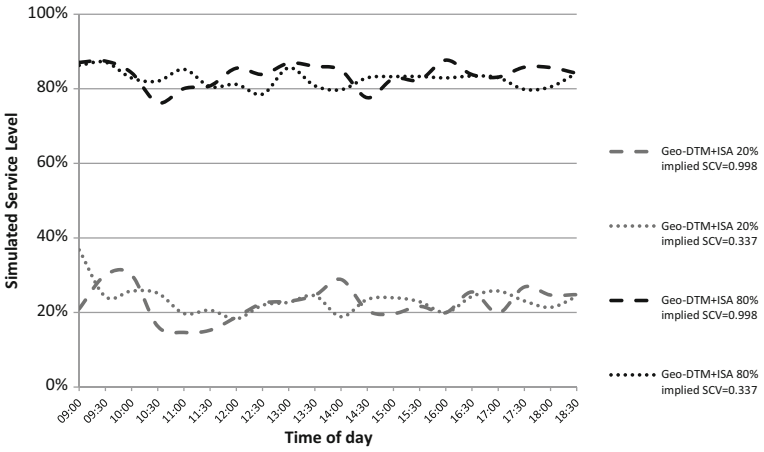

Fig. 11 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA pre-emptive policy with different SCVs (Beta [2,5] service, TSF within 0 s, pre-emptive policy)

The benefits of matching the SCV of service time in this way has been investigated for various Beta and LogNormal distributions with SCV < 1.0, for both pre-emptive and exhaustive policies, see Chen (2014) for further details. Figure 11 shows typical results, comparing the service levels from Geo-DTM+ISA with SCV = 0.998 (i.e. matching the exponential distribution, as in Fig. 9) with ones obtained using Geo-DTM+ISA with SCV = 0.337 (i.e. matching the Beta[2,5] distribution). This has significantly improved the stability and the consistency with which the service levels achieved adhere to the specified service target lines throughout the day.

Because it is not possible to produce Geometric distributions with SCV > 1.0, it is not possible to use Geo-DTM+ISA to produce better results than those shown in Fig. 10. Using the traditional DTM would make this possible, but at some computational cost.

4.4 $M_t/G/s_t + G$

To investigate the extent to which the performance of Geo-DTM+ISA deteriorates when used on cases with non-exponential abandonment times, it has been tested on systems of the type $M_t/G/s_t+G$, where the abandonment times take Exponential, Erlang and Lognormal distributions which vary in SCV and in the shape of their left-hand tails, as in Whitt (2005). The four distributions used are Exponential (SCV = 1), Erlang 2 (SCV = 0.5), Lognormal (SCV = 1) and Lognormal (SCV = 4), and their CDFs are shown in Fig. 12. With many abandonments likely to take place within the first 60–90 s, it is of particular interest that the Exponential and Lognormal (SCV = 4) distributions are close in the shape of their left-hand tails, as are the Erlang 2 and Lognormal (SCV = 1) distributions.

Figures 13, 14 and 15 show the simulated performance of systems of the form $M_t/G/s_t+G$ that have been staffed using Geo-DTM+ISA. For example, in Fig. 13 (which introduces Erlang 2 distributions of service and abandonment times) and in Fig. 14 (which introduces Lognormal (SCV = 1) distributions of service and abandonment times), when the TSF is high there is very little deterioration in the performance of Geo-DTM+ISA, and service levels only rarely drop a little below the target of 80 %. For more congested systems with TSF = 20 %, the introduction of Erlang 2 or Lognormal (SCV = 1) service times again causes little deterioration in performance, as was previously observed in Sect. 4.3 for systems without abandonments.

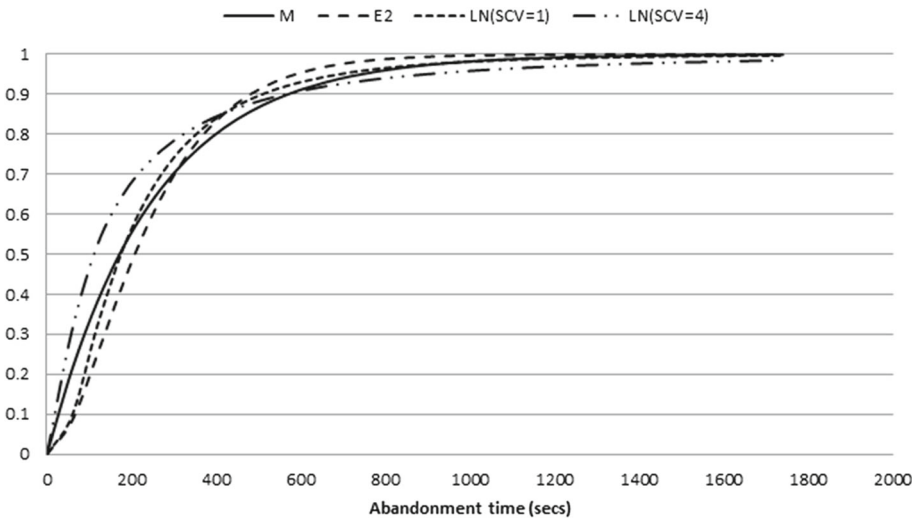

Fig. 12 CDFs of the four abandonment time distributions: Exponential (SCV = 1), Erlang 2 (SCV = 0.5), Lognormal (SCV = 1) and Lognormal (SCV = 4)

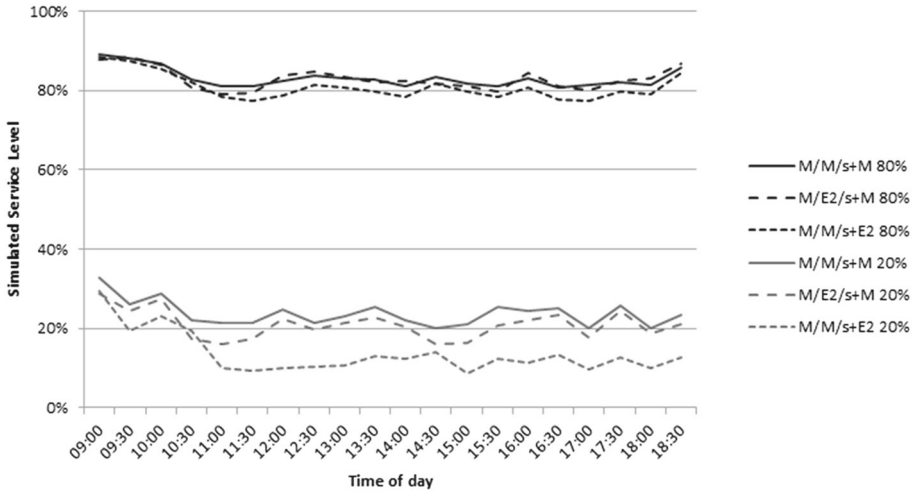

Fig. 13 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA for Erlang 2 distributed service and abandonment times (TSFs within 0 s, pre-emptive policy)

However the impact of introducing Erlang 2 or Lognormal (SCV = 1) abandonment times is much greater, as was seen in some of Whitt's (2005) examples for steady-state behaviour of large call centres. Comparing Figs. 13 and 14 it can also be observed that the impacts of Erlang 2 and Lognormal (SCV = 1) abandonment times are quite similar, which is attributable to their quite similar left-hand tails, and is despite their quite different SCVs.

In contrast Fig. 15 shows a bigger impact of the Lognormal (SCV = 4) service time compared to the Lognormal (SCV = 4) abandonment time. In this case we see from Figure 12 that the left-hand tails of Exponential and Lognormal (SCV = 4) distributions are very similar, and so there is little impact on performance when the distribution of abandonment time is

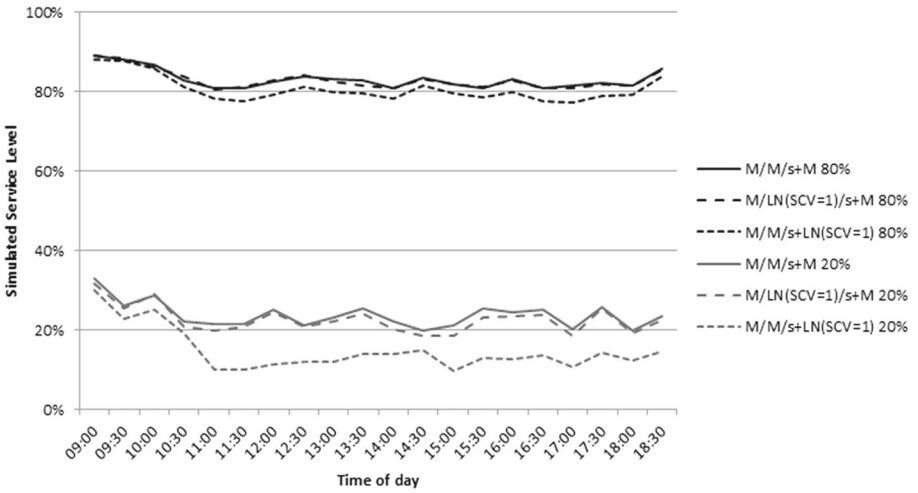

Fig. 14 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA for Log-normal (SCV = 1) distributed service and abandonment times (TSFs within 0 s, pre-emptive policy)

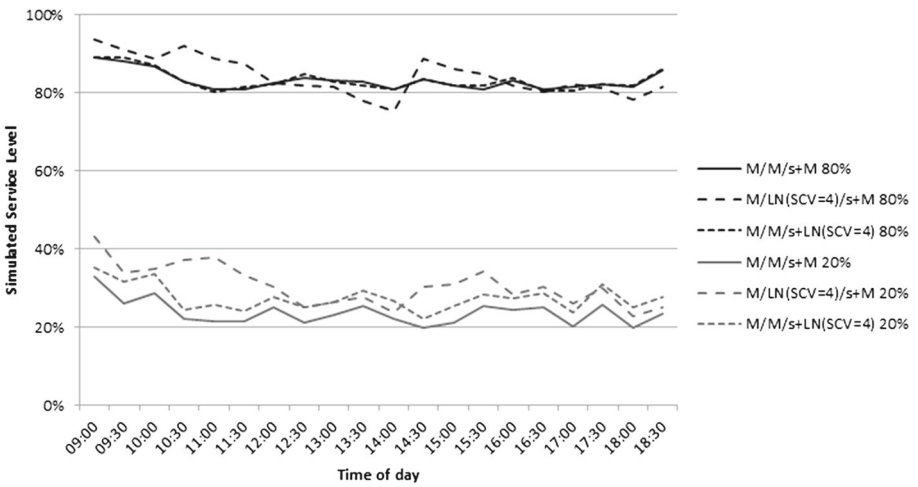

Fig. 15 Simulated service levels based on the staffing level patterns generated by Geo-DTM+ISA for Log-normal (SCV = 4) distributed service and abandonment times (TSFs within 0 s, pre-emptive policy)

changed. However their SCVs are very different, and hence contribute to the relatively large impact when the distribution of service time is changed.

Whitt (2005) shows that for large steady-state systems *M/M/s+M(n)* models can serve as good approximation for $M/G/s+G$ systems, and hence an area for further work is to investigate whether $M_t/Geo/s{+}Geo(n)$ models can serve as good approximations for $M_t/G/s{+}G$ systems.

4.5 Computational experience

The runtime of Geo-DTM+ISA is considerably faster than that of simulation based methods. The extent of the runtime speed differential depends on maximum number of customers

Table 1 Comparison of the computation time (in seconds) for Geo-DTM and simulations (2000 runs), with different maximum numbers allowed in the system

Maximum number in the system (L)	Geo-DTM (180 steps, implied SCV = 0.190) (s)	Geo-DTM (60,000 steps, implied SCV = 0.998) (s)	Simulation (2000 runs) (s)
100	2	18	10,325
200	12	71	11,162
400	34	420	12,327
800	89	2458	12,883

allowed in the system (L) and the stepsize that is chosen when fitting the Geometric distribution of service time. Table 1 compares the times taken to estimate the performance of the call centre cases over the 10 h day using Geo-DTM+ISA and using simulation. Note that although $L \leq 200$ was sufficiently accurate for the medium sized call centres considered here, running the models with larger L was undertaken to show the computational requirements if larger L were required.

Hence for the call centre scenarios considered in this paper there is over a 150-fold benefit in runtime if the default stepsize is used when fitting the Geometric distribution, and it could be over 900-fold if the service time has a small SCV.

5 Discussion and conclusions

The iterative staffing algorithm described in Sect. 3, based on the one described in Feldman et al. (2008), worked well for our wide range of systems. It differs from their algorithm as it does not require very short staffing intervals, and it introduces a binary search once the initial bounds s_2 and s_3 have been found. Our empirical results also include systems of the form $M_t/G/s_t$ and support the contention that the approach is robust and widely applicable.

Results shown in Sects. 4.1 and 4.2 are based on extensive test cases of the form $M_t/M/s_t$ and $M_t/M/s_t + M$ in which Geo-DTM+ISA was used to generate staffing functions for some typical call centre scenarios with time-dependent arrivals in order to achieve a range of time-stable service level targets in a wide range of experimental settings: time service factors representing*efficiency-driven* and *quality-driven*; with or without abandonments; and exhaustive or pre-emptive policies. In all cases Geo-DTM+ISA was found to consistently achieve service levels which were stable and adhered to the specified service target levels throughout the day.

Furthermore, Geo-DTM+ISA consistently outperforms approaches that are widely used in industry and their lagged counterparts, i.e. it consistently outperforms Erlang-C+PSA and Erlang-C+LPSA for systems of the form $M_t/M/s_t$, and Erlang-A+PSA and Erlang-A+LPSA for systems of the form $M_t/M/s_t+M$. The extent to which Geo-DTM+ISA outperforms the other methods depends mainly on the performance of those methods. The methods using lagged PSA are generally better than those just using PSA, and the errors introduced by both PSA and lagged PSA methods are generally smaller when there are only low levels of congestion, for example in *quality driven* systems and in systems with high levels of abandonments.

Results in Sect. 4.3 for systems of the form $M_t/G/s_t$, where the service time distributions are chosen with squared coefficients of variation (SCVs) ranging from 0.077 to 2.0 and pre-emptive and exhaustive polices, show that Geo-DTM+ISA continues to outperform the PSA and lagged PSA methods. However it nolonger consistently adheres to the target service levels, with examples which sometimes deviate below target service levels (when SCV $<$ 1.0), and sometimes deviate unnecessarily above target service levels (when SCV $>$ 1.0).

For systems of the form $M_t/G/s_t$ it is also shown that when the service time SCV < 1.0 it is possible to choose a time step size so that the selected Geometric distribution has the same SCV. In this case the performance of Geo-DTM+ISA improves significantly, and is close to its levels of stability and consistency previously observed for systems of the form $M_t/M/s_t$. For the case of SCV > 1.0 no better Geometric distribution can be found than that used to match an exponential distribution. In the case of SCV > 1.0 better performance could in theory be achieved using the traditional DTM approach, but this would be at some computational cost, and is an area yet to be researched.

Results in Sect. 4.4 for systems of the form $M_t/G/s_t+G$, where the abandonment time distributions are chosen with squared coefficients of variation (SCVs) ranging from 0.5 to 4.0, and with very different shaped left-hand tails, show circumstances under which the performance of Geo-DTM+ISA deteriorates. An area for further work is to investigate whether $M_t/Geo/s+Geo(n)$ models can serve as good approximations for $M_t/G/s+G$ systems in these circumstances.

The methodology based on analytical models is computationally more efficient than one based on simulation methods. For the medium sized call centre test cases used in this research the saving is between 150-fold and 900-fold. However it is also interesting to note that simulation becomes more competitive as the call centre size increases, with the benefits of Geo-DTM+ISA dropping to between 5-fold and 140-fold for the biggest test cases.

In common with Feldman et al. (2008), in order to evaluate Geo-DTM+ISA we have used test cases in which it is assumed that the time-dependent arrival rates are known. However, as noted in Koole (2013), in a real call centre environment the time-dependent arrival rates will typically be forecasts, which will be imperfect and thus subject to forecast errors. Hence another aspect of this research, to be reported elsewhere, investigates the effects of forecast errors on the performance of queueing models for staffing.

Appendix 1

Forward recurrence algorithm for the $M(t)/Geom/S$ model

a) Initialise all probability distributions and set starting conditions.

```
π₀(0) = 1
for n = 1,2 …, L
 πₙ(0) = 0
next n
```

b) Iterating forward through times (t). Take each state (n) in turn, and run through all the possible numbers of departures (d) and possible numbers of arrivals (r). Each case leads to a resultant state at time ($t + 1$) and an associated probability, which are accumulated to give the final probabilities of each possible state at time ($t + 1$).

```
Iterate forwards through time (t).

for t = 0,1,2, ….. T
   Consider each state of possible number of customers in the system (n).

   for n = 0,1,2 …., L

   Calculate the number of customers in service between t and t+1.
   Set c = min(n,s)

      Consider each possible number of customers completing service between t
and t+1

      for d = 0,1, …c

      Calculate the probability of d departures: S_d|c based on (2)
         Consider each possible number of arrivals at the end of the interval

         for r = 0,1,2,…L

         Calculate the probability of arrivals r: V_r, see (3)
         Calculate resulting state n', i.e.

         Set n' = min (L, n − d + r) as in (4).

Calculate the associated probability of the resulting state.
         π_n(t+1) := π_n × S_d|c × V_r
         {where the operator ':=' indicates the accumulation of probabilities}

         next r

      next d

   next n

next t
```

Appendix 2

Forward recurrence algorithm for the $M(t)/Geom/S+Geom$ model

Having first set the initial conditions as in "Appendix 1".

```
Iterate forwards through time (t).
for t = 0,1,2, ….. T
   Consider each state of possible number of customers in the system (n).
   for n = 0,1,2 …., L
   Calculate the number of customers in service between t and t+1.
   Set c = min(n,s)
   Calculate the number of customers in the queue between t and t+1
   Set q = min(0,n − s)
      Consider each possible number of abandonments (a)from the queue between t and
t+1
      for a = 0, 1, …. q
      Calculate A_{a|q}(t) using equation(6)
        Consider each possible number of customers completing service between t and
t+1 (d)
        for d = 0,1, …c
         Calculate the probability of departure d:  S_{d|c} based on (2)
           Consider each possible number of arrivals at the end of t
           for r = 0,1,2,…L
           Calculate the probability of arrivals r: V_r
           Calculate resulting state n'
           n' = min (L, n − a − d + r)
           Calculate the associated probability of the resulting state.
           π_n(t + 1) := π_n × A_{a|q} × S_{d|c} × V_r
           {where the operator ':=' indicates the accumulation of probabilities}
           next r
        next d
     next a
   next n
next t
```

References

Abdalla, N., & Boucherie, R. J. (2002). Blocking probabilities in mobile communications networks with time-varying rates and redialling subscribers. *Annals of Operations Research*, *112*, 15–34.

Bekker, R., & de Bruin, A. M. (2010). Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, *178*, 45–65.

Chassioti, E., & Worthington, D. J. (2004). A new model for call centre queue management. *The Journal of the Operational Research Society*, *55*(12), 1352–1357.

Chassioti, E., Worthington, D. J., & Glazebrook, K. (2014). Effects of state-dependent balking on multi-server non-stationary queueing systems. *Journal of the Operational Research Society*, *65*, 278–290.

Chen, X. (2014). Combining forecasting and queueing models for call centre staffing, *Ph.D. Thesis*, University of Lancaster, UK.

Dafermos, S., & Neuts, M. F. (1971). A single server queue in discrete time. *Cahiers du Centre de Recherche Operationnelle*, *13*(1), 23–40.

Defraeye, M., & Van Nieuwenhuyse, I. (2013). Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decision Support Systems*, *54*, 1558–1567.

Feldman, Z., Mandelbaum, A., Massey, W. A., & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science*, *54*(2), 324–338.

Galliher, H. P., & Wheeler, R. C. (1958). Nonstationary queuing probabilities for landing congestion of aircraft. *Operations Research*, *6*(2), 264–275.
Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, *5*(2), 79–141.
Green, L., & Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, *37*(1), 84–97.
Green, L., Kolesar, P., & Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, *16*(1), 13–39.
Griffiths, J., Holland, W., & Williams, J. (1991). Estimation of queues at the channel tunnel. *Journal of the Operational Research Society*, *42*(5), 365–373.
Ingolfsson, A. (2005). *Modeling the M(t)/M/s(t) queue with an exhaustive discipline*. Working paper, Department of Finance and Management Science, School of Business, University of Alberta, Edmonton, Alberta, Canada.
Izady, N., & Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research*, *219*(3), 531–540.
Jennings, O. B., Mandelbaum, A., Massey, W. A., & Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science*, *42*(10), 1383–1394.
Koole, G. (2013). *Call center optimization*. Amsterdam, The Netherlands: MG Books.
Neuts, M. F. (1973). The single server queue in discrete time numerical analysis I. *Naval Research Logistics Quarterly*, *20*, 297–304.
Wall, A. D., & Worthington, D. J. (2007). Time dependent analysis of virtual waiting time behaviour in discrete time queues. *European Journal of Operational Research*, *178*(2), 482–499.
Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science*, *51*(2), 221–235.
Worthington, D., & Wall, A. (1999). Using the discrete time modelling approach to evaluate the time-dependent behaviour of queueing systems. *The Journal of the Operational Research Society*, *50*(8), 777–788.